Submit for publication for ISIT'94

Adaptive Source Coding Schemes for Geometrically Distributed Integer Alphabets *

Kar-Ming Cheung

Jet Propulsion Laboratory
Pasadena, CA 91109 USA

ABSTRACT:

In this article we revisit Gallager and van Voorhis optimal source coding scheme for geometrically distributed non-negative integer alphabets, and show that the various subcodes in the popular Rice algorithm can be derived from the Gallager and van Voorhis code. Next we modify and generalize the Gallager and van Voorhis code for two-sided geometrically distributed integer alphabets (positive and negative), which are typical input samples to the back-red entropy coding stage of lossless predictive coding schemes and lossy transform coding schemes. We develop an adaptive coding scheme, and show that this adaptive coding scheme has low implementation complexity. We present some theoretical and experimental results.

* The research described in this paper was carried out by Jet Propulsion Laboratory, California Institute of Technology, under a contract with National Aeronautics and Space Administration

Adaptive Source Coding Schemes for Geometrically Distributed Integer Alphabets *

EXTENDED ABSTRACT

1. Introduction

In this article we revisit Gallager and van Voorhis optimal source coding scheme for geometrically distributed non-negative integer alphabets [1]

$$p_{G1}(i) = (1 - \theta)\theta^i \quad \forall\ \theta \geq 0, \tag{1}$$

where $\theta = 1 - r(0)$, and $r(0)$ is the fraction of zeros in the sample set. We show that the various subcodes in the popular Rice algorithm can be derived from the Gallager and van Voorhis code, Next we modify and generalize the Gallager and van Voorhis code for 2-sided geometrically distributed integer alphabets (positive and negative), which have the following distribution

$$p_{G2}(i) = \frac{1-\theta}{1+\theta}\theta^{|i|} \quad \forall\ i, \tag{2}$$

where $\theta = \frac{1-r(0)}{1+r(0)}$, and $r(0)$ is the fraction of zeros in the sample set. The 2-sided geometrically distributed integer alphabets are typical inputs to the back-end entropy coding stage of lossless predictive coding schemes and lossy transform coding schemes. We develop an adaptive coding scheme which has low implementation complexity, and we present some theoretical and experimental results of this scheme.

II. Relationship Between the one-sided Gallager van Voorhis Code and the Rice Code

Gallager and Van Voorhis presented an optimal binary prefix code for the set of geometrically distributed nonnegative integers [1]. Here we call this code the Gallager-van Voorhis-Huffman-1 (GVH1) code. Let l be the integer satisfying

$$\theta^l + \theta^{l+1} \leq 1 < \theta^l + \theta^{l-1}, \tag{3}$$

where $\theta = 1 - r(0)$ as defined in (1). It is easy to see that for any θ, $0 < \theta < 1$, there is a unique positive integer l satisfying (3). Let a non-negative number i be represented by $i = lj + r$ where $j = \lfloor i/l \rfloor$, the integer part of i/l, and $r = [i]$ 1110111. Gallager and Van Voorhis showed that an optimal code for the non-negative integers is the concatenation of a unary code which is used to encode j, and a Huffman code which is used to encode ?", $0 < r \leq l-1$.

Rice developed a predictive lossless coding scheme [2] that consists of two separate stages: the front-end pre-processor is a predictor followed by a symbol mapper, while the second part performs adaptive entropy coding. The first stage takes the difference between the actual values and the predicted values and maps the differences, positive or negative, to a sequence of 11011- negative integer numbers. The second stage encodes the sequence by adaptively selecting the best of several easily implemented variable length coding algorithms for non-negative integers. Using Rice's notation in [2], it was shown in [3] that the various variable length codes constructed by concatenating the fundamental sequence code Ψ_1 and the split-sample codes $\Psi_{1,k}$ are optimal Huffman codes for data sources that have Laplacian distributions. In this article we further show that the optimal Huffman codes in the Rice algorithm are actually particular GVH1 codes that correspond to those l's which are powers of two. Omitting the mathematical details, we show that the fundamental sequence code Ψ_1 is equivalent to the GVH1 code for $l = 1$, and the split-sample codes $\Psi_{1,k}$ is equivalent to the GVH1 code for $l = 2^k$.

* The research described in this paper was carried out by Jet Propulsion Laboratory, California Institute of Technology, under a contract with National Aeronautics and Space Administration

III. Efficient Coding Based on the 2-sided Geometric Model

Constructing an optimal prefix code, say by using the Huffman algorithm, is quite a complex operation in hardware. We developed a class of near-optimal prefix codes to encode data (e.g. differentials of waveform data and image data) with probability distributions that resemble the 2-sided geometric models as introduced in the previous section [4]. The construction of this prefix code is simple. For most well-behaved data, frequency(i) $\approx$ frequency($-i$) for $i = 1, 2, \ldots$. Thus in order to construct a code for both the positive and negative values, we use the GVH codes for the non-negative integers. An additional bit is appended to each codeword, except the codewords representing 0, to indicate whether integer i or integer $-i$ is sent. We call this code the Gallager-van Voorhis-Huffman-2 code.

Based on the above code construction, we evaluate the performance of the GVH2 codes, and give closed form analytic expressions as a function of θ for the redundancy r_2, the mean codelength $\bar{l}_2$, and the entropy $H(X_2)$ of the 2-sided integer geometric distribution, where X_2 is the discrete random variable corresponding to the 2-sided geometric source. We show that $\bar{l}_2$ is given by

$$\begin{aligned}\bar{l}_2 &= \frac{2}{1+\theta}\left(\lfloor\log_2(l)\rfloor + 1 + \frac{\theta^k}{1-\theta^l}\right) + 1 - \frac{1-\theta}{1+\theta} - \frac{1-\theta}{1+\theta}\left(1 + \lfloor\log_2(l)\rfloor\right) \\ &= 1 + \lfloor\log_2(l)\rfloor + \frac{2}{1+\theta}\left(\theta + \frac{\theta^k}{1-\theta^l}\right)\end{aligned} \tag{4}$$

We also show that the entropy of the 2-sided geometric source can be written as

$$\begin{aligned}H(X_2) &= \sum_{i=-\infty}^{i=\infty} p_2(i) l_2(i) \\ &= \log_2\left(\frac{1+\theta}{1-\theta}\right) - \frac{2\theta\log_2(\theta)}{(1+\theta)(1-\theta)}\end{aligned} \tag{5}$$

Hence we write down a closed form expression for the redundancy of our coding scheme as a function of θ and l, namely,

$$\begin{aligned}\bar{r}_2 &= \bar{l}_2 - H(X_2) \\ &= 1 + \lfloor\log_2(l)\rfloor + \frac{2}{1+\theta}\left(\theta + \frac{\theta^k}{1-\theta^l}\right) - \log_2\left(\frac{1+\theta}{1-\theta}\right) + \frac{2\theta\log_2(\theta)}{(1+\theta)(1-\theta)}\end{aligned} \tag{6}$$

We find the value of l which minimises $\bar{r}_2$ for given θ by minimising the terms in $\bar{r}_2$ which depend on l, namely

$$f(l) = \lfloor\log_2(l)\rfloor + \frac{2}{1+\theta}\left(\frac{\theta^k}{1-\theta^l}\right) \tag{7}$$

We find the optimal l values (over all ranges of θ of interest) by direct search, and the results will be presented in the conference.

IV. An Adaptive Coding Scheme Based on 2-Sided Geometric Distribution

In this section we describe an adaptive entropy coding scheme that was developed for the Galileo Low Gain Antenna Mission [6]. This adaptive lossless data compression sheme is differential-pulse-code-modulation based (DPCM based), and uses a Huffman coding strategy similar to the one used to compress the DC differentials of the JPEG [7] and ICT [6] [8] compression schemes. We develop three Huffman codebooks that are based 011 the 2-sided geometry model: one for low-activity data (l = 1), one for medium-activity data (l = 2), and one for high-activity data (l = 4). The data are first partitioned into blocks of fixed length (e.g. 16 samples per block). The first sample of each block is used as a reference point and is not coded. And for the remaining samples, the differences between adjcent samples are calculated. The encoder then computes the number of bits that are required to compress the Mock using each of the predefined codel inks, and C]1OOSCS the codebook that gives the best compression. If all codel rooks give data-expansion, the block is sent unencoded. Each block is preceded by a 2-bit tag: 00 for the low-activity codel rook, ()1 for the medium-activity codebook, 10 for the high-activity codebook, and 11 for 110 compression. Simulation results on various data sources will be given at the conference.

References:

[1] R. Gallager and D. van Voorhis, "Optimal Source Codes for Geometrically Distributed Integer Alphabets," *IEEE Trans. Inform. Theory*, vol. IT-21, March 1975.

[2] R. Rice, "Some Practical Universal Noiseless Coding Techniques," Part I and II, JPL Publications 79-22 and 83-17, March 1979.

[3] P. Yeh, R. Rice, and W. Miller, "On the Optimality of a Universal Noiseless Coder," *Proceedings of the AIAA Computing in Aerospace 9 Conference*, San Diego, October 1993.

[4] K. Cheung and P. Symth, "A High-Speed Distortionless Predictive Image Compress Scheme," TDA Progress Report 42-101 vol. January-March, 1990, Jet Propulsion Laboratory, Pasadena, CA.

[5] P. Smyth, "Entropy-based bounds on the redundancy of prefix codes," presented at the IEEE International Symposium on Information Theory, San Diego, 1990.

[6] K. Cheung and K. Tong, "Proposed Data Compression Schemes for the Galileo S-Band Contingency Mission," *Proceedings of the 1993 Space Earth Science Data Compression Workshop*, Snowbird, Utah, April 2, 1993.

[7] W. Pennebaker and J. Mitchell, "JPEG Still Image Data Compression Standard", New York, van Nostrand Reinhold, 1993.

[8] W. Cham, "Development of Integer Cosine Transform by the Principle of Dyadic Symmetry," *IEE Proceedings, Vol 136, August 1989.*